A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

Thank you for agreeing to review this paper for Annals of Botany. The Annals of Botany aims to be among the very top of plant science journals and as we receive over 1000 submissions every year we need to be very selective in deciding which papers we can publish. In making your assessment of the manuscript's suitability for publication in the journal please consider the following points.

Scientific Scope

Annals of Botany welcomes papers in all areas of plant science. Papers may address questions at any level of biological organization ranging from molecular through cells and organs, to whole organisms, species, communities and ecosystems. Its scope extends to all flowering and non-flowering taxa, and to evolutionary and pathology research. Many questions are addressed using comparative studies, genetics, genomics, molecular tools, and modeling.

To merit publication in Annals of Botany, contributions should be substantial, concise, written in clear English and combine originality of content with potential general interest.

- We want to publish papers where our reviewers are enthusiastic about the science: is this a paper that you would keep for reference, or pass on to your colleagues? If the answer is “no” then please enter a low priority score when you submit your report.

- We want to publish papers with novel and original content that move the subject forward, not papers that report incremental advances or findings that are already well known in other species. Please consider this when you enter a score for originality when you submit your report.

Notes on categories of papers:

All review-type articles should be **novel, rigorous, substantial and “make a difference” to plant science**. The purpose is to summarise, clearly and succinctly, the “cutting edge” of the subject and how future research would best be directed. Reviews should be relevant to a broad audience and all should have a **strong conclusion and illustrations** including diagrams.

- *Primary Research* articles should report on original research relevant to the scope of the journal, demonstrating an important advance in the subject area, and the results should be clearly presented, novel and supported by appropriate experimental approaches. The Introduction should clearly set the context for the work and the Discussion should demonstrate the importance of the results within that context. Concise speculation, models and hypotheses are encouraged, but must be informed by the results and by the authors' expert knowledge of the subject.

- *Reviews* should place the subject in context, add significantly to previous reviews in the subject area and moving forward research in the subject area. Reviews should be selective, including the most important and best, up-to-date, references, not a blow-by-blow and exhaustive listing.

- *Research in Context* should combine a review/overview of a subject area with original research, often leading to new ideas or models; they present a hybrid of review and research. Typically a Research in Context article contains an extended Introduction that provides a general overview of the topic before incorporating new research results with a Discussion proposing general models and the impact of the research.

- *Viewpoints* are shorter reviews, presenting clear, concise and logical arguments supporting the authors' opinions, and in doing so help to stimulate discussions within the topic.

- *Botanical Briefings* are concise, perhaps more specialised reviews and usually cover topical issues, maybe involving some controversy.

Article type: Original Article

Title:

A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

Kangshan Mao[1, †, *], Markus Ruhsam[2, †], Yazhen Ma[1], Sean W. Graham[3], Jianquan Liu[1], Philip Thomas[2], Richard I. Milne[4], Peter M. Hollingsworth[2]

Affiliation:

[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065, Sichuan, P. R. China

[2]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK

[3]Department of Botany, University of British Columbia, Vancouver, V6T 1Z4, Canada

[4]Institute of Molecular Plant Science, School of Biological Science, University of Edinburgh, Edinburgh, EH9 3BF, UK

[†] These authors contributed equally to this work

Running title: A transcriptome-based phylogeny for junipers and cypresses

*For correspondence: maokangshan@scu.edu.cn or maokangshan@163.com

ABSTRACT

Background and Aims Rapid evolutionary divergence and reticulate evolution may result in phylogenetic relationships that are difficult to resolve using small nucleotide sequence datasets. Next-generation sequencing methods can generate larger datasets that are better suited to solving these puzzles. One major and long-standing controversy in conifers concerns generic relationships within the subfamily Cupressoideae (105 species, ~1/6 of all conifers) of Cupressaceae, in particular the relationship between *Juniperus*, *Cupressus* and the *Hesperocyparis-Callitropsis-Xanthocyparis* (HCX) clade. Here we attempt to resolve this question using transcriptome-derived data.

Methods Transcriptome sequences of 20 species from Cupressoideae were collected. Using MarkerMiner, single copy nuclear (SCN) genes were extracted. These were applied to estimate phylogenies based on concatenated data, species trees, and a phylogenetic network. We further examined the effect of alternative backbone topologies on downstream analyses, including biogeographic inference and dating analysis.

Results Based on the 73 SCN genes (>200,000 bp total alignment length) we considered, all tree-building methods lent strong support for the relationship (HCX, (*Juniperus*, *Cupressus*)); however, strongly supported conflicts among individual gene trees was also detected. Molecular dating suggests that these three lineages shared a most recent common ancestor ~60 Mya, and that *Juniperus* and *Cupressus* diverged ~56 Mya. Ancestral area reconstructions (AARs) suggest an Asian origin for the entire clade, with subsequent dispersal to North America, Europe and Africa.

Conclusions Our analysis of SCN genes resolves a controversial phylogenetic relationship in the Cupressoideae, a major clade of conifers, and suggests that rapid evolutionary divergence and incomplete lineage sorting likely acted together as the source for conflicting phylogenetic inferences between gene trees and between our robust results and recently published studies. Our updated backbone topology has not substantially altered molecular dating estimates relative

to previous studies, however application of the latest AAR approaches has yielded a clearer picture of the biogeographic history of Cupressoideae.

Key words: single-copy nuclear genes, transcriptome, Cupressoideae, *Hesperocyparis*, *Cupressus*, *Juniperus*, *Xanthocyparis*, *Callitropsis*

INTRODUCTION

It can be challenging to accurately reconstruct deep phylogenetic relationships within groups that experienced rapid evolutionary divergence, incomplete lineage sorting and/or reticulate evolution, especially with small datasets (Maddison, 1997; Dunn *et al*., 2008; Jian *et al*., 2008; Zeng *et al*., 2014; Ruhsam *et al.,* 2015). Rapid evolutionary divergence may lead to short internodal distances and soft polytomies (Weisrock *et al*., 2005; Whitfield & Lockhart, 2007; Jian *et al*., 2008; Pyron *et al*., 2014; Leaché *et al*., 2016). In addition, incomplete lineage sorting, which involves mis-sorting of ancestral polymorphisms relative to the species tree, or reticulate evolution, which involves the combination or transmission of genetic material between divergent evolutionary lineages due to hybridization and introgression, may both cause inaccurate or conflicting species-tree inference (Beiko *et al*., 2008; Sun *et al*., 2015).

Next-generation sequencing approaches, which generate large amounts of DNA sequence data from throughout the genome, are transforming phylogenetic inference (e.g. Dunn *et al*., 2008; Lee *et al*., 2011; Faircloth *et al*., 2012; Zeng *et al*., 2014). This is especially true where rapid evolutionary events resulted in few fixed substitutions between divergent species, yielding gene trees that are usually unresolved with respect to the true species tree, when only a few loci are used (Whitfield & Lockhart, 2007). A larger amount of sequence data is likely to capture such species-specific substitutions, potentially resulting in improved phylogenetic resolution (Jian *et al*., 2008; Zeng *et al*., 2014). In the case of incomplete lineage sorting, many independent gene trees from throughout the genome can be used to estimate a credible species tree by reconciling genealogical discordance between loci (Edwards, 2009; Lemmon and Lemmon, 2013). Therefore, phylogenetic estimation of species trees based on genomic datasets might resolve branches that were poorly supported in smaller datasets (Rokas *et al*., 2003; Dunn *et al*., 2008). For example, phylogenetic analyses using as few as 29 and 59 low-copy nuclear genes have resulted in well-resolved deep phylogenetic estimates for ferns (Rothfels *et al*., 2015) and flowering plants (Zeng *et al*., 2014), respectively.

Two main methods have recently been proposed to construct species trees from large datasets (Liu *et al.*, 2009a, 2009b, 2015). One method uses the multiple-species coalescent model as implemented in the program *BEAST (Heled and Drummond, 2010), which estimates gene trees and the species tree at the same time. However, this method is computationally intensive (Edwards *et al*., 2007; Pyron *et al*., 2014), and may result in poor convergence if the dataset is large (O'Neill *et al*., 2013). The other method uses a two-step approach when estimating species trees. In the first step gene trees are generated using software such as RaxML (Stamatakis *et al*., 2014), and in the second step they are summarized under the coalescent model as implemented in the software MP-EST (Liu *et al*., 2010), STAR (Liu *et al*., 2009a). This method reduces computation time considerably when compared to analyses based on the multiple species coalescent model (Liu *et al*., 2009b). In addition, a recently developed two-step approach, ASTRAL-II (Mirarab *et al.*, 2015; Mayyari and Mirarab, 2016), has been shown to run much faster and to be less sensitive than MP-EST to the effects of gene tree errors, when estimating a species tree based on large dataset (e.g. hundreds of taxa and thousands of genes). The accuracy of ASTRAL remains high when a small number of genes is adopted and a moderate level of incomplete lineage sorting is assumed, whereas its local posterior probabilities of quartet branches are conservative; this leads to very few false positives that have high support, at the cost of missing some true positives (Mayyari and Mirarab, 2016).

Cupressaceae, also known as the cypress family, contains more than 160 species in 32 genera, of which 17 are monotypic (Farjon, 2005; Mao *et al*., 2012; Yang *et al*., 2012; Wang & Ran, 2014; Adams 2014). They occur in many different habitats on all continents except Antarctica (Farjon, 2005). Cupressoideae, which contains more than 100 species in 13 genera, is the largest of the seven subfamilies of Cupressaceae (Gadek *et al*., 2000; Mao *et al*., 2012; Yang *et al*., 2012). This subfamily occurs throughout the Northern Hemisphere and contains many ecologically important and dominant species especially in mountainous and arid or semi-arid regions (Farjon, 2005; Adams 2014). It contains many economically important timber species (e.g. *Calocedrus*, *Chamaecyparis*, *Cupressus* and *Thuja*) and ornamental trees (e.g.

Chamaecyparis, *Juniperus*, *Platycladus* and *Thuja*) (Farjon, 2005). Phylogenetic analyses suggest that this subfamily is monophyletic (Gadek *et al*., 2000; Mao *et al*., 2012; Yang *et al*., 2012) and comprises four clades (Gadek *et al*., 2000; Little 2006; Mao *et al*., 2012; Yang *et al*., 2012) which have been treated as separate tribes by some authors (Gadek *et al*., 2000). However, taxonomic treatment at the generic level and inter-generic relationships within the subfamily remains controversial (Little *et al*., 2004; Little 2006; Mill & Farjon, 2006; Rushforth, 2007; Adams *et al*., 2009; Christenhusz *et al*., 2011; Dörken *et al*., 2017), especially for *Cupressus* sensu lato (s.l.), which comprises 30 species (Little, 2006; Christenhusz *et al*., 2011; Dörken *et al*., 2017). *Cupressus* s.l. may be divided into four genera: *Cupressus* sensu stricto (s.s.) and *Xanthocyparis* s.s. in the Old World, and *Hesperocyparis* and *Callitropsis* s.s. in the New World (Adams *et al*., 2009; Mao *et al*., 2010; Christenhusz *et al*., 2011) (see Table 1 for a summary of taxonomic treatment history). Henceforth, if not stated otherwise, "*Cupressus",* "*Xanthocyparis",* and "*Callitropsis*" refer to *Cupressus* s.s., *Xanthocyparis* s.s. and *Callitropsis* s.s., respectively. Although the monophyly of *Cupressus* and the *Hesperocyparis*-*Callitropsis*-*Xanthocyparis* clade (the HCX clade; Mao *et al*., 2010) is well defined (Little *et al*., 2004; Little 2006; Mao *et al*., 2010, 2012; Yang *et al*., 2012), the phylogenetic relationship between *Cupressus*, the HCX clade and *Juniperus* remains uncertain. All possible phylogenetic topologies among these three clades have been supported by different studies with different datasets and analyses, as follows: (*Cupressus*, (*Juniperus*, HCX)) topology was recovered by Xiang & Li (2005), Adams *et al*. (2009) and Terry & Adams (2015); (*Juniperus*, (*Cupressus*, HCX)) topology by Mao *et al.* (2010); (HCX, (*Cupressus*, *Juniperus*)) topology by Little (2006) and Yang *et al*. (2012); and a trichotomy (HCX, *Cupressus*, *Juniperus*) by Mao *et al*. (2012). From here on, these topologies are referred to as Cu(HCX,Ju), Ju(Cu,HCX), HCX(Cu,Ju) and (HCX,Cu,Ju), respectively, for simplicity. A recent phylogenomic study based on the whole plastid genomes of 22 species of Cupressaceae and accounting for long branch attraction (e.g., Felsenstein, 1978; Hendy and Penny, 1989) supported the Ju(Cu,HCX) topology (Qu *et al*., 2017). However, all of these studies either used no more than four bi-parentally inherited nuclear loci (e.g. Little 2006; Adams

et al., 2009) or plastid DNA (ptDNA), the latter of which, despite the use of nine (Mao *et al*., 2010), 11 ptDNA regions (Terry & Adams, 2015) or even the whole plastid genome (Qu *et al*., 2017), can be considered to be a single locus due to its lack of recombination.

The aim of the current study is to resolve this long-standing controversy and to reconstruct the phylogenetic relationship between *Cupressus*, *Juniperus* and the HCX clade based on a number of single or low copy nuclear loci from transcriptome data using 17 species representing major lineages within these three clades, plus three outgroups. Specifically, we investigate (a) the evolutionary relationship between the three major lineages using a phylotranscriptomic approach, (b) compare and explain the discordance and agreement between the current species tree topology and phylogenetic topologies that were gained in previous studies, and characterize the impact of different topologies of the three major lineages on (c) molecular dating of this group and (d) the inference of its biogeographic history.

MATERIALS AND METHODS

Provenance of samples

Fresh leaf samples from a total of 18 species (including outgroup species *Microbiota decussata*) were collected for transcriptome sequencing. Fourteen samples were collected from the living collection of the Royal Botanic Garden Edinburgh (RBGE), three were collected in the field in Yunnan, China (*Cupressus duclouxiana*) and Xizang, China (*Cupressus gigantea* and *Juniperus microsperma*), and one (*Cupressus funebris*) was a cultivated individual from the campus of Sichuan University, Chengdu, China (Table 2). Additionally, we used transcriptome data for three outgroup species (*Calocedrus decurrens*, *M. decussata*, *Thuja plicata*) from the one thousand transcriptome project (‘1000 plant project,’ 1KP) (Table 2). All species were represented by a single accession apart from *M. decussata* (n=2).

Transcriptome sequencing, assembly, and alignment

Transcriptomes were either generated in Edinburgh/UK (RBGE, Table 2) or Chengdu/China (SZ, Table 2) apart from three downloaded from the 1KP project (http://www.onekp.com/samples/list.php; labelled as '1kp' in Table 2). RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma-Aldrich, St. Louis, Missouri, USA) following protocol A with a few minor modifications (2-3 times the amount of lysis buffer, 750 µl binding buffer and three final washes). Library preparation and sequencing was outsourced to Edinburgh Genomics (Edinburgh, UK) and Novogene (Beijing, China) for RBGE and SZ samples, respectively (Table 2). Transcriptomes were sequenced on Illumina HiSeq platforms generating 2 × 100 bp paired-end reads. Raw reads were prepared for assembly using Trimmomatic (Bolger *et al.*, 2014) with the parameters 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36' and cutadapt (Martin 2011) to remove adapters and low quality sequences. Reads for each taxon were then assembled into contigs with SOAPdenovo- trans (Xie *et al.*, 2014) using SOAPdenovo-Trans-31mer with '-K 29 -L 100'. The programme Cd-hit (Li & Godzik 2006), software for clustering and comparing protein or nucleotide sequences, was used to retrieve only unique contigs from the SOAPdenovo-trans analysis with the command cd-hit-est and default values. The output of Cd-hit was then fed into MarkerMiner v1.0 (Chamala *et al.*, 2015) with parameters '-singleCopyReference Athaliana -minTranscriptLen 900'. MarkerMiner identifies and aligns putative orthologous single or low copy nuclear genes in a set of transcriptome assemblies, using a reciprocal BLAST search against a reference database (Chamala *et al.*, 2015). The alignments of genes included for further analyses were visually checked with minimal editing and trimming either side of the sequence where missing sites accounted for more than half of all available taxa. In subsequent analyses data from the two *Microbiota* accessions (Table 2) were amalgamated to represent one sample in order to minimise the amount of missing data for that species.

Phylogenetic analyses

Alignments of putative single copy loci from MarkerMiner v1.0 (Chamala *et al.*, 2015) were used to compile two sets of data, the first comprising individual genes, in which each locus is treated independently, and the second a concatenated dataset in which all chosen loci were combined into one 'super locus'. First, we used three conventional methods, MP, ML and Bayesian Inference to infer phylogenetic trees based on the concatenated dataset: analyses based on such dataset could lead to species-tree mis-inference if there is sufficient conflict between gene trees, but these concatenation-based methods often recovers the same tree that other species-tree estimation methods recover (e.g. Wickett *et al*., 2014) and is commonly done to compare to other species-tree estimation methods, e.g. MP-EST (Liu *et al.*, 2010), STAR (Liu *et al.*, 2009a), ASTRAL (Mirarab *et al*., 2014). Maximum parsimony analysis (MP) was performed using PAUP*4.0b10 (Swofford 2003), with gaps treated as missing data and polymorphic states as uncertain. A 'branch and bound' search with MulTrees on was carried out for both datasets. Branch support was estimated via bootstrapping with 1000 bootstrap replicates using heuristic searches (Felsenstein, 1985). We also used RAxML v8 (Stamatakis 2014) to estimate a maximum likelihood (ML) tree and ML bootstrap values, by applying the parameters '-f a -m GTRGAMMA -p 12345 -x 12345 -# 1000' where the GTRGAMMA model and 1000 bootstrap replicates were applied (see RAxML manual for detailed parameter settings). A Bayesian inference analysis (BI) was also performed using MrBayes v. 3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) with the GTR+I+G model, which was selected using MrModelTest v. 2.3 (Nylander 2004) under the Akaike Information Criterion. The analysis was run for 2 million generations with four MCMC chains in two independent parallel analyses, with one tree sampled every 500 generations. The average standard deviation of split frequencies was 0.00000 at the end of the run. TRACER v1.5 (Rambaut & Drummond 2009) was used to assess the quality of the MCMC simulations and suggested a high degree of convergence between runs. The effective sample size values (ESS), that is the number of effectively independent draws from the posterior, were >500 for all parameters, indicating that sufficient sampling occurred.

From the individual gene dataset, we constructed individual ML gene trees for each locus using the software RAxML v8 (Stamatakis 2014) applying the parameters as for the concatenated data set. The topology of each gene tree was then manually examined, looking in particular for well-supported alternative relationships that might indicate gene-tree conflict. Then, based on these gene trees, we generated a species tree based on the multispecies coalescent model in ASTRAL 5.6.1 (Mirarab *et al*., 2014; Mayyari and Mirarab, 2016), which estimates species trees from unrooted gene trees, and maximizes the number of quartet trees shared between the gene trees and the species tree. ASTRAL-II estimates branch lengths for internal branches (not terminal branches) in coalescent units, and branch support values measure the support for a quadripartition (the four clusters around a branch) and not the bipartition, as is commonly done. The species tree was fully annotated using "-t 4" option, which calculates the measurements for each branch, including quartet support (q), total number of quartet trees in all the gene trees (f), and the local posterior probabilities (pp) for the main topology and the first, second alternatives, total number of quartets defined around each branch (QC), and the effective number of genes for the branch (EN).

Conservative pp scores cause some true positives to be overlooked in ASTRAL (Mayyari and Mirarab, 2016); moreover, average positive branch rates, which represent the proportion of the estimated species tree in which a certain branch is successfully recovered, may be lower in ASTRAL than in STAR and MP-EST (Liu *et al*., 2015). Therefore, we also conducted STAR and MP-EST analyses based on gene trees to reduce the chance of missing any true positives, and to improve the average positive branch rates. Hence the rooted 'best tree' RAxML output for each gene plus bootstrap values for each gene tree using 1000 replicates was then uploaded to 'The Species TRee Analysis Webserver' STRAW (Shaw *et al.*, 2013) to estimate the species tree using STAR (Liu *et al.*, 2009a) and MP-EST (Liu *et al.*, 2010). Both programmes apply the multispecies coalescent model (Rannala & Yang 2003) to obtain estimates of the species tree from gene trees. STAR (Liu *et al.*, 2009a) uses the average ranks of coalescences, whereas

MP-EST (Liu *et al.*, 2010) uses a pseudo-likelihood function of the species tree, and both of them generate bootstrap support values using nonparametric bootstrap techniques (Liu *et al.*, 2009a, 2010). Both methods are based on summary statistics calculated across all gene trees, with the effect that a small number of genes that significantly deviate from the coalescent model will have relatively little effect on the ability to accurately infer the species tree.

Because there was some well-supported conflict among gene trees (see Results), we conducted two additional analyses to investigate this further. First, we applied MulRF (Chaudhary *et al.*, 2013, 2015) to estimate the best species tree, i.e. the one that minimizes the overall Robinson-Foulds (RF) distance between each candidate species tree and the individual gene trees. This software is also able to calculate the MulRF score of a given tree topology, which is the RF distance between this given tree and all gene trees. In a soft polytomy where relationship among three clades are difficult to resolve, this function may be used to compare the compatibility of each of the three dichotomy candidate species trees with all gene trees.

Finally, we used the NeighborNet method implemented in SplitsTree 4.11.3 (Huson and Bryant, 2006) to reconstruct phylogenetic networks based on the concatenated alignment of all 73 nuclear genes. For distance calculations, we excluded insertions/deletions (indels) and used the K2P model (Kimura, 1980). The relative robustness of the clades was estimated by performing 1000 bootstrap replicates, and a confidence network was generated with a 95% threshold (Huson and Bryant, 2006). This analysis can summarize how homoplasy that might include hybridization or incomplete lineage sorting might have affected the phylogenetic reconstruction.

Molecular dating

To investigate the impact of topological differences on the evolutionary divergence timescale in Cupressoideae, we conducted molecular dating analyses. We tried to adopt the eight fossil calibration points as in Mao *et al*. (2010), but only three of them could be used for the dating of our 20-taxon data set, whereas the remaining five could only be attached to apparently deeper

nodes, relative to those in the phylogeny of Mao *et al.* (2010). As too few calibration points and/or assigning fossils to deeper nodes (due to sparse sampling) has been shown to bias the estimates of node ages (e.g. Linder *et al*., 2005; Mao *et al*., 2012; Wang & Mao, 2016), we adopted a hybrid strategy to reconstruct the evolutionary divergence timescale of Cupressoideae. Hence dating was carried out on our previous ptDNA dataset comprising nine ptDNA fragments from 84 species (Mao *et al*., 2010), but with the relationship between the three main clades constrained to the topology from the current study based on transcriptome data (see below). The original ptDNA dataset comprising 92 accessions was slightly reduced by removing multiple accessions of six species, resulting in a final dataset of 86 accessions representing 84 species in Cupressoideae (referred to as '86-accession data set' from here on). Three parallel molecular dating analyses were carried out, one constraining to the HCX(Cu,Ju) topology, another constraining to the Cu(HCX,Ju) topology, and the third was unconstrained, allowing it to retain the Ju(Cu,HCX) topology from Mao *et al*. (2010). We adopted eight calibration fossils from Mao *et al*. (2010), seven of which were used as minimum age constraints with uniform priors, and one was set as a fixed age constraint with a normal prior (see Table 1 in Mao *et al*., 2010 for details).

BEAST version 1.8.0 (Drummond & Rambaut, 2007) was used to simultaneously estimate topology, substitution rates and node ages by employing a Bayesian MCMC chain. BEAST parameter settings, including fossil calibration settings, were all the same as in Mao *et al*. (2010), except that two independent MCMC analyses of 100 000 000 generations were conducted, sampled every 2 000 generations, with 20% burn-in. The program Tracer 1.5.1 (Rambaut & Drummond, 2007) was employed to check effective sample size, and the program TreeAnnotator 1.8.0 (part of the BEAST 1.8.0 package) was used to summarize the output results. Finally, a tree with ages for each node and their 95% highest posterior density intervals (95%HPD), was displayed and formatted in FigTree 1.3.1 (Rambaut, 2008).

Ancestral area reconstruction

We conducted an ancestral area reconstruction using the BioGeoBEARS packages as implemented in RASP 4.0 (Yu *et al*., 2015). Four operational geographic areas (A: North America, B: Africa, C: Asia, D: Europe), were defined for our analyses, following those in Mao *et al*. (2010). A total of 400 trees, which were resampled from the output trees of the BEAST analysis, and the BEAST summary tree, were imported into RASP 4.0, along with the distribution information of each species. BioGeoBEARS allows the testing of six models (DIVALIKE, DIVALIKE+J, DEC, DEC+j, BAYAREALIKE and BAYAREALIKE+J) (Matzke 2013, 2014). Of the six models, the DIVALIKE model (Ronquist, 1997) is an event-based approach that adopts a simple biogeographic model; it does not consider general area relationships or branch lengths of the input tree, and it applies different costs to vicariance, duplication, dispersal, and extinction to construct ancestral distributions (Ronquist, 1997; Yu *et al*., 2015). The DEC model (Ree and Smith, 2008) allows dispersal, extinction, and cladogenesis as fundamental processes, accommodates differing dispersal probabilities among areas across different time-periods, and can integrate branch lengths, divergence times, and geological information (Ree and Smith, 2008; Yu *et al*., 2015). In contrast to the former two models, which accept only bifurcating trees, the BAYAREALIKE model (Landis *et al*., 2013) allows polytomies. It considers distribution area to be a 'trait' of a species, and hence reconstructs ancestral 'traits' using Bayesian inference; furthermore, it does not define the dispersal rate, constrain the maximum number of areas at each node, or exclude widespread and unlikely ancestral areas before analysis (Landis *et al*., 2013; Yu *et al*., 2015). The other three models with the '+J' suffix (i.e., DIVALIKE+J, DEC+j, BAYAREALIKE+J) allow founder-event speciation, in contrast to the three original models (Matzke, 2014).

We conducted model testing and two models (the best and the second-best model, as given in the Results section) were employed to reconstruct the ancestral area for every node in the phylogeny based on 100 trees that were randomly selected from 400 BEAST trees. At most two areas were allowed for any node in any tree. An among-area dispersal probability matrix, which is the same as in Mao *et al*. (2012), was coded to define different dispersal probabilities in five

time periods, 0-5 Mya, 5-30 Mya, 30-45 Mya, 45-70 Mya and 70-115 Mya. The ancestral area reconstruction results were optimized in the Treeview window of the RASP program.

RESULTS

Phylogenetic analyses

A total of 581 putative single or low copy genes were detected by MarkerMiner. About 50% of these were shared by five or fewer samples. To minimise the impact of missing data on the ability to confidently resolve phylogenetic relationships, we only included genes that did not have more than two ingroup and two outgroup taxa missing, for further phylogenetic analyses. This resulted in 73 putatively single or low copy genes yielding an alignment of 208,484 nuclear base pairs for 20 taxa (Supplementary Table S1). The DNA sequences have been deposited in NCBI GenBank (accession numbers shown in Table S2).

For six of the 73 genes chosen by MarkerMiner (Chamala *et al*., 2015), a secondary transcript that passed the BLAST filtering process was reported for one of the included taxa (Supplementary Table S1). These secondary transcripts, defined as the one of the two from a given taxon that received a lower BLAST score, may represent splice isoforms, putative paralogs, or partially assembled transcripts (Chamala *et al*., 2015). However, as removing the taxon from the particular alignment for which a secondary transcript was detected did not change the phylogenetic results in any way, for these species (data not shown), the transcript for each taxon with the higher BLAST score was included in all analyses.

In this matrix consisting of 73 genes, 183,022 (87.8%) characters were constant, 16,405 (7.9%) were variable but parsimony-uninformative and 9,057 (4.3%) were parsimony-informative. The same topology was retrieved regardless of the bifurcate tree-building method used (including MP, ML, BI, MP-EST, STAR, ASTRAL-II, MulRF), with an HCX clade (comprising a monophyletic *Hesperocyparis*, plus *Callitropsis* and

353 *Xanthocyparis*) as the sister group of a clade consisting of *Cupressus* (monophyletic) and
354 *Juniperus* (also monophyletic) (Figs 1, 2). Branch support was high for the MP, ML and BI
355 analyses of the concatenated data set and for the MP-EST, STAR and ASTRAL-II analyses
356 using a coalescent approach. Individual trees produced by Bayesian analysis and the three
357 coalescent approaches were identical. The MP tree topology differed from these in only one
358 respect: here (*J. procera* (*J. indica, J. microsperma*)) was sister to (*J. flaccida, J. scopulorum*)
359 (Supplementary Fig. S1), whereas in the other analyses *J. procera* was sister to ((*J. indica*, *J.*
360 *microsperma*) (*J. flaccida, J. scopulorum*)) (Figs 1, 2). However, two branches concerning this
361 relationship were weakly or moderately supported by the bootstrap analysis in the MP analysis
362 (BS=55% and 76%, respectively, Fig. S1). The *Juniperus* clade was also the only ingroup clade
363 where some of the internal relationships did not receive maximum branch support; this was true
364 for all six analyses methods used (MP, BI, ML, MP-EST, STAR and ASTRAL-II; Figs 1, 2).
365 The HCX(Cu,Ju) topology of our species tree based on single nuclear genes conflicts with the
366 Ju(Cu,HCX) topology based on the ptDNA data from Mao *et al*. (2010) (Supplementary Fig. S2).
367 Quartet support analyses in ASTRAL-II suggest that the HCX(Cu,Ju), Ju(Cu,HCX) and
368 (Cu(HCX,Ju) topologies are supported by 54.16%, 24.24% and 21.60% of the gene trees,
369 respectively (Fig. 2). A manual check of gene trees (Fig. S3) which were generated in RAXML
370 using maximum likelihood bootstrapping (MLBS) indicated similar proportions of gene trees
371 supporting these three topologies (i.e., most supporting the HCX(Cu,Ju) topology, which is
372 equivalent to the *Cupressus*-*Juniperus* sister topology in Table S1), except that a few MLBS
373 gene trees are unresolved (Table S3). We also calculated the MulRF score (the total
374 Robinson-Foulds topological distance of all 73 gene trees against the candidate species tree) for
375 each of the above three topologies concerning *Cupressus*, *Juniperus* and HCX (other
376 relationships remain the same). The HCX(Cu,Ju), Ju(Cu,HCX) and Cu(HCX,Ju) topologies
377 received MulRF scores of 744 (= closest and therefore favored), 786 and 790 (= furthest),
378 respectively.

Finally, NeighborNet analyses provide 100% bootstrap support for the quartet branch that links (*Juniperus*, *Cupressus*) and (HCX, outgroups), although the length of this branch is relatively short (Fig. 3A). Very few strongly supported relationships that might have suggested hybridization or incomplete lineage sorting (bootstrap support >95%) were recovered in the NeighborNet confidence network; these were mainly found within *Juniperus*, but also once within *Cupressus* (the branch leading to (*C. gigantea*, *C. duclouxiana*)), at the basal position of the HCX clade, and among the three outgroups (Fig. 3B).

Molecular dating

The BEAST analysis based on the two phylogenetic topologies, HCX(Cu,Ju) and Ju(Cu,HCX), yielded effective sample sizes that were well above 200 (> 800) for branch lengths, topology and clade posteriors and all other relevant parameters, indicating adequate sampling of the posterior distribution. However, the BEAST analysis based on the Cu(HCX,Ju) topology failed, despite being identical to other analyses in all but enforcement of topology, because every one of >20 attempts returned an error message that the log likelihood of the initial tree is negative Infinity.

Based on the HCX(Cu,Ju) topology that was supported by single-copy nuclear (SCN) genes, we estimate that the most recent common ancestor (MRCA) of *Cupressus*, *Juniperus* and the HCX clade diverged from the MRCA of *Platycladus*, *Microbiota*, *Calocedrus* and *Tetraclinis* 81.06 Mya [70.50-90.40](from here on, shown in square brackets are the 95% HPD range of age estimation), the HCX clade diverged from the MRCA of *Cupressus* and *Juniperus* 59.80 Mya [48.45-71.74], and *Juniperus* diverged from *Cupressus* 56.33 [45.30-67.95] Mya. The crown ages of the HCX clade, *Cupressus* and *Juniperus* were estimated to be 37.45 Mya [23.54-52.30], 28.73 Mya [16.65-42.15], and 41.34 Mya [29.99-44.63], respectively (Fig. 4, Table 3).

Based on the Ju(Cu,HCX) topology that was supported by ptDNA tree, age estimation for all nodes overlapped with the above estimation (see Table 3) except that the HCX clade diverged from *Cupressus* 54.09 Mya (95% HPD: 41.29-67.03). A comparison of age estimation of major nodes in BEAST analyses based on each of the above two topologies are shown in Table 3.

406

407 **Ancestral area reconstruction**

408 Model tests in the BioGeoBEARS package, based on either the HCX(Cu,Ju) or the Ju(Cu,HCX)
409 topology, suggested that DIVALIKE+J is the best-performing model (AICc_wt values:
410 HCX(Cu,Ju) topology = 0.65, Ju(Cu,HCX) topology = 0.60), whereas DEC+J model is the
411 second-best model (AICc_wt values: HCX(Cu,Ju) topology = 0.33, Ju(Cu,HCX) topology =
412 0.39).

413 Based on the HCX(Cu,Ju) topology, the DIVALIKE+J model and the 86-accession data set,
414 *Cupressus*, *Juniperus* and the HCX clade share a common ancestor whose ancestral distribution
415 area is probably Asia (ca. 0.96), whereas *Cupressus* and *Juniperus* shared a common ancestor
416 whose ancestral distribution area is likely to be Asia (ca. 0.82) or less likely Europe (ca. 0.16).
417 The ancestral area for the MRCA of the HCX clade is inferred to be Asia (ca. 0.54), North
418 America (ca. 0.33) or a combination of these two (ca. 0.13). Within this clade, the common
419 ancestor of all New World cypresses (*Callitropsis* plus *Hesperocyparis*) most likely migrated to
420 and diversified in North America later (Fig. 5). The ancestral area for *Cupressus* is probably Asia
421 (ca. 0.99), and *Cupressus sempervirens* and its close allies dispersed to Europe around the middle
422 Miocene (Fig. 5). Furthermore, the ancestral area for *Juniperus* is inferred to be Europe (ca.
423 0.72), or possibly Asia (ca. 0.23). The common ancestor of sect. *Juniperus* was inferred to be in
424 Europe (ca. 0.56), Asia (ca. 0.36) or a combination of both, whereas that of sect. *Sabina* was
425 probably in Asia (ca. 0.65) and possibly in Europe (ca. 0.25) or Africa (ca. 0.05); overall,
426 *Juniperus* is most likely to have diversified within Eurasia, with three separate dispersal events
427 to North America, and one to Africa. BioGeoBEARS analysis based on the DEC+J model
428 yielded highly similar results (not shown), especially concerning major nodes in the phylogeny.

429 Based on the Ju(Cu,HCX) topology, either the DIVALIKE+J or the DEC+J model and the
430 86-accession data set, the ancestral area for nearly all nodes are highly similar to the HCX(Cu,Ju)
431 topology, except for the node of the common ancestor of *Cupressus* and the HCX clade, which

does not exist in the HCX(Cu,Ju) topology. The ancestral area for this node in the Ju(Cu,HCX) topology is likely to be Asia (DIVALIKE+J: ca. 0.97; DEC+J: 0.95) (Supplementary Fig. S4).

DISCUSSION

Rapid evolutionary divergence and inference of phylogenetic relationships among the three major clades in Cupressoideae

The main aim of this paper is to resolve and explain the long-standing controversy of generic and inter-generic relationships between the three major lineages in Cupressoideae, *Cupressus*, the *Hesperocyparis-Callitropsis-Xanthocyparis* (HCX) clade, and *Juniperus*. Our phylogenetic analyses using maximum parsimony (MP), maximum likelihood (ML), Bayesian inference (BI) analyses of concatenated data and species tree analyses (MP-EST, STAR, and ASTRAL), based on 73 putative single copy nuclear genes totaling more than 200,000 base pairs, all show a maximally supported sister relationship of *Cupressus* and *Juniperus*, and that their common ancestor is sister to the HCX clade (Figs 1, 2). Although only weakly or moderately supported, the Ju(Cu,HCX) topology based on ptDNA (Supplementary Fig. S2; Mao *et al*., 2010) conflicts with the HCX(Cu,Ju) topology here, as well as several published phylogenies (Xiang & Li, 2005; Little *et al*., 2006; Adams *et al*., 2009; Yang *et al*., 2012; Terry & Adams, 2015). This incongruence may have been caused by incomplete lineage sorting due to rapid evolutionary divergence and/or hybridization and introgression between the three clades during their early evolutionary history. Yang *et al*. (2012) constructed a reticulate network using two nuclear loci, and because relationships of the three subclades were incongruent among different datasets, suggested that *Cupressus* "might have originated through hybridization between *Juniperus* and the ancestor of *Hesperocyparis–Callitropsis–Xanthocyparis*" (Yang *et al*., 2012; p462). However, although hybridization and introgression during earlier history is a possibility, the main cause of the phylogenetic pattern among the three clades appears to be a combination of rapid evolutionary divergence and incomplete lineage sorting.

First, all phylogenetic analyses we conducted based on 73 loci support the HCX(Cu,Ju) topology. As we have shown above, the species tree constructed using MP-EST, STAR, ASTRAL or trees built using MP, ML, BI based on concatenated data show 100% support for the HCX(Cu,Ju) topology. The species tree estimate from MulRF also supports the HCX(Cu,Ju) topology: in particular, the RF distance between all gene trees to the HCX(Cu,Ju) topology is closer than to either the Ju(Cu,HCX) topology or the Cu(HCX,Ju) topology. In addition, Neighbor-Net tree based on concatenated dataset also support the HCX(Cu,Ju) topology with 100% bootstrap support (Fig. 3A), and the "reticulate" pattern among the three clades that Yang *et al.* (2012) reported was not detected (Fig. 3).

Second, the gene tree topology frequency we found here may fit better with incomplete lineage sorting as an explanation of conflicting gene trees. The maximum support value for nodes in the species tree does not necessarily mean that there is no conflict between the 73 gene trees. In our ASTRAL analyses, for example, the HCX(Cu,Ju), Ju(Cu,HCX) and Cu(HCX,Ju) topologies, received quartet support values of 54.16%, 24.24% and 21.60%, respectively (equivalent to 39.54, 17.69, 15.77 gene trees, respectively). We further checked the MLBS tree for each of the 73 genes and found that 38, 15, 14 gene trees support the above three topologies, respectively; if only MLBS values above 70% are considered (corresponding to a moderately well supported branch), then 31, 11 and 7 gene trees supported the three topologies, respectively. If conflict between gene trees is caused by incomplete lineage sorting, which is always a close companion of rapid evolutionary divergence (e.g., Maddison, 1997), then we would expect one high frequency topology and two lower frequency topologies. Conversely, if the conflict between gene trees is caused by hybridization and introgression (e.g. the hypothesis that *Cupressus* is a hybrid clade between *Juniperus* and HCX that Yang *et al*. (2012) have put forward), one might expect two major (and equivalent) frequency gene tree topologies (e.g., if the branch was a result of hybrid speciation) or some other set of frequencies (e.g., if a subset of the genome introgressed at this point). To conclude, the pattern of gene tree topology frequencies we found

above is more consistent with the scenario of incomplete lineage sorting than the hybridization and introgression scenario.

Third, the internode branch lengths between the three clades are consistently short in both our species tree (Fig. 2) and trees based on concatenated data (Fig. 1), and the molecular dating suggest that the interval between the MRCA of HCX-*Juniperus*-*Cupressus* (~59.8 Mya) and the MRCA of *Juniperus*-*Cupressus* (~56.3 Mya) is also relatively short (~3.5 Myr), consistent with rapid evolutionary divergence (and presumably a substantial chance of retaining some conflicting ancestral polymorphisms, as documented for our individual gene trees, Fig. S3). This has also been the case in previous phylogenies of Cupressoideae. For example, using the whole plastid genome the inferred internode branch length is short, regardless of whether the HCX(Cu,Ju) or Ju(Cu,HCX) topology is recovered (Qu *et al*., 2017), and based on six ptDNA regions the phylogenetic relationship of these three clades remained unresolved (Mao *et al*., 2012). However, there is one exception to the pattern, which is that the internode branch length between the MRCA of the three clades and the MRCA of *Cupressus* and *Juniperus* is relatively long based on the nuclear gene NEEDLY (Yang *et al*., 2012).

Thus, although we cannot exclude reticulate evolution in shaping the current phylogenetic pattern of the three clades within Cupressoideae, rapid evolutionary divergence better explains the pattern we found. This inference is different from another case in this subfamily where reticulate evolution is clearly indicated among *Thuja* species (Peng and Dan, 2008).

Transcriptomic data provide strong support for a four-genus taxonomic treatment in *Cupressus* s.l.

Previous studies suggested four possible phylogenetic topologies concerning these three clades. Phylogenetic analyses based on either three or six ptDNA markers show that these three clades are part of a trichotomy (Little *et al*., 2006; Mao *et al*., 2012), whereas nine ptDNA markers provide moderate support for the Ju(Cu,HCX) topology (Mao *et al*., 2010). A recent study based

on whole plastid genomes supported the clustering of *Juniperus* and *Cupressus*, while a filtered dataset, which was meant to reduce or elucidate long branch artifacts, supported the clustering of *Cupressus* and the HCX clade (Qu *et al*., 2017). The Cu(HCX,Ju) topology was supported by a series of studies: based on nrITS region alone (Xiang & Li, 2005), a combined dataset that included nrITS, two ptDNA markers and 56 morphological characters (Little *et al*., 2004); a combined dataset that included one ptDNA region and three nuclear regions (nrITS, ABI3, 4CL) (Adams *et al*., 2009); and a combined dataset that included 11 ptDNA regions and two nuclear regions (nrITS and NEEDLY) (Terry & Adams *et al*., 2015). Phylogenetic analyses based on a single nuclear region, NEEDLY (MP support value: 100%), and a combined dataset that included three ptDNA regions, two nuclear regions (ITS, NEEDLY) and 88 organismal characters (MP support value: 100%; Little, 2006) supported a HCX(Cu,Ju) topology, in agreement with our SCN results (Fig. 1).

One important implication of our results is that *Cupressus* s.l. is paraphyletic, and should be divided into four genera (see Fig. 1 and Table 1 for a clarification of taxon names). Nearly all published molecular phylogenetic analyses support the monophyly of both *Cupressus* s.s. and the HCX clade, yet the sister relationship between them is rarely supported (Mao *et al*., 2010). Hence, Little (2006) proposed to call the HCX clade *Callitropsis* s.l., where *Xanthocyparis* s.l. was merged into *Callitropsis* s.l., yet such a treatment is not universally accepted. Considering a proposal of Mill and Farjon (2006) to conserve the genus name *Xanthocyparis*, which was ratified by the International Botanical Congress in 2011 (Barrie, 2011), and that *Xanthocyparis* s.l. is not monophyletic, Adams *et al*. (2009) proposed to place all New World cypresses (*Cupressus* sensu Farjon species in North America) in the new genus *Hesperocyparis* and keep both *Xanthocyparis* s.s. and *Callitropsis* s.s. as monotypic genera. Our results support this, showing that both *Cupressus* s.l. and *Xanthocyparis* s.l. are paraphyletic, while each of *Cupressus*, the HCX clade, and *Hesperocyparis* is monophyletic. Hence, our nuclear-based results strongly support the division of *Cupressus* s.l. into four genera: *Cupressus*, *Hesperocyparis*, *Xanthocyparis* s.s. and *Callitropsis* s.s. (Adams *et al*., 2009; Mao *et al*., 2010)

and rejects both the combination of these four genera under *Cupressus* s.l. (Christenhusz *et al*., 2011; Table 1) and the combination of *Xanthocyparis* s.s. and *Callitropsis* s.s. under either *Xanthocyparis* s.l. or *Callitropsis* sensu Little (2004). Our data, as well as many others (e.g. Little, 2006; Mao *et al.*, 2010), would also be consistent with combining *Xanthocyparis* s.s. and *Callitropsis* s.s. and *Hesperocyparis* under *Callitropsis* s.l., yet *Hesperocyparis* is morphologically distinct enough to deserve recognition as a distinct genus (Adams *et al*., 2009).

An updated evolutionary divergence timescale and biogeographic history of *Cupressus*, the HCX clade and *Juniperus*

Rerunning the molecular dating on the ptDNA dataset from Mao *et al.* (2010) while constraining it with the nuclear species tree (HCX(Cu,Ju)) topology of our results, suggests that the split between the *Cupressus*-*Juniperus* clade and the HCX clade occurred (48.45-) 59.80 (-71.74) Mya, with a split of the former clade into *Cupressus* and *Juniperus* happening only 3.47 Myr later, (45.30-) 56.33 (-67.95) Mya. Comparing this to the Ju(Cu,HCX) topology that was supported by ptDNA data (Mao *et al*., 2010), the only difference is that *Juniperus* diverges first (47.54-) 59.44 (-71.24) Mya, followed by the divergence of *Cupressus* from HCX (41.29-) 54.09 (-67.03) Mya, in the Ju(Cu,HCX) topology. All other nodes occur in both topologies and differ in age between topologies by no more than 1.04 Myr, a difference dwarfed by HPD error ranges (Fig. 4, Table 3). This indicates that, in our case, a single topological difference, even in the deep nodes in a phylogeny, had very limited effect on node age estimates. A possible reason for this may be that this particular topological difference did not alter the phylogenetic position of fossil calibration points, and barely affected the total length between any given node and the root of the tree (Sauquet *et al*., 2012; Wang & Mao, 2016).

We also reran the ancestral area reconstruction analyses for both the HCX(Cu,Ju) and Ju(Cu,HCX) topologies using BioGeoBEARS, and four parallel analyses were conducted for each of the two topologies based on two different models (DIVALIKE+J, DEC+J). Apart from the MRCA of *Cupressus* and *Juniperus*, and the MRCA of *Cupressus* and the HCX clade, that

566 are specific to the HCX(Cu,Ju) and Ju(Cu,HCX) topologies, respectively, the relative probability
567 of the ancestral area for all other nodes in all four parallel analyses are highly similar. We
568 therefore discuss the reconstructed biogeographic history of the HCX clade, *Cupressus* and
569 *Juniperus* based on the HCX(Cu,Ju) topology and the best model (DIVALIKE+J model). Our
570 ancestral area reconstruction (AAR) analysis inferred that both the MRCA of the HCX clade,
571 *Cupressus* and *Juniperus*, and the MRCA of *Cupressus* and *Juniperus*, most likely originated in
572 Asia. Likewise, the HCX clade most likely originated in Asia and then dispersed once to North
573 America and diversified there (Fig. 5). This fits a pattern of directional migration from the
574 northwest to the southeast in North America in New World Cypresses (*Callitropsis* and
575 *Hesperocyparis*), which may have been caused by climate cooling and aridification in the latter
576 half of the Cenozoic (Terry *et al*., 2016). *Cupressus* probably originated in Asia, and then
577 dispersed to Europe (and northern Africa) around the middle Miocene (Fig. 5). The genus
578 *Juniperus* and sect. *Juniperus* most likely originated in Europe, whereas sect. *Sabina* originated
579 in Asia; three independent migrations from Eurasia to North America and one migration from
580 Eurasia to Africa were inferred (Fig. 5).
581 Comparing these results to the previous AAR analysis based on S-DIVA (Mao *et al*., 2010),
582 the AAR analysis based on BioGeoBEARS yielded a clearer resolution, especially concerning
583 the ancestral area of the MRCA of the HCX clade, the MRCA of *Juniperus*, the MRCA of
584 *Juniperus* sect. *Juniperus* and sect. *Caryocedrus*, the MRCA of *Juniperus* sect. *Juniperus*, and
585 the MRCA of Clade I (*Juniperus pseudosabina* plus all Himalayan/Qinghai-Tibet Plateau
586 species except *J. microsperma* and *J. gausenii*) and Clade II (serrate-leaved junipers of North
587 America) (Mao *et al*., 2010). BioGeoBEARS tends to infer a single area as the ancestral area,
588 whereas the S-DIVA usually infers the combination of two disjunct areas as the ancestral area.
589 The integration of dispersal probability among areas during different time periods in the past,
590 and the use of a model test to seek the best-performing model are likely to have improved the
591 resolution of AAR in BioGeoBEARS compared to S-DIVA.
592

Conclusion

Phylogenetic relationships among *Cupressus*, *Hesperocyparis-Callitropsis-Xanthocyparis* (HCX) and *Juniperus* have been a contentious issue since the discovery of the Golden Vietnamese Cypress, *Xanthocyparis vietnamensis*. Our species tree based on 73 nuclear loci yielded 100% support for a (HCX, (*Cupressus*, *Juniperus*)) topology which is in agreement with previous phylogenies based on two nuclear loci (LEAFY and NEEDLY; Yang *et al*., 2012) and a combined dataset including both morphological characters and molecular dataset (Little, 2006), but contradicts many others. This indicates that *Cupressus* s.l. (Christenhusz *et al*., 2011; Table 1) is paraphyletic, and can be considered instead as two monophyletic genera, *Cupressus* (s.s.) and *Hesperocyparis*, and two monotypic genera, *Callitropsis* (s.s). and *Xanthocyparis* (s.s.). Rapid evolutionary divergence and incomplete lineage sorting may have been the major cause for the minor conflicts observed among gene trees. Molecular dating based on the nuclear species tree (HCX(Cu,Ju)) topology suggests that the three clades underwent two evolutionary splits in a time period as short as ca. 3.47 Myr. The split between *Cupressus*+*Juniperus* and the HCX occurred ca. 59.80 Mya (95%HPD: 48.45-71.74 Mya), and the split between *Cupressus* and *Juniperus* occurred ca. 56.33 Mya (95%HPD: 45.30-67.95 Mya). Ancestral area reconstruction analyses suggest that the MRCA of *Juniperus* probably occurred in Europe, whereas the MRCAs of HCX, *Cupressus*, *Cupressus*+*Juniperus*, and HCX+*Cupressus*+*Juniperus* all most likely occurred in Asia. Therefore, the common ancestor of these three clades most likely originated in Asia and then diversified and dispersed to Europe, North America and Africa. Our study shows that combining low copy nuclear genes collected using next generation sequencing and coalescent-based species tree estimation methods is a powerful approach that provides more refined phylogenetic estimates of deep nodes in conifer phylogeny that were controversial based on small datasets.

Acknowledgements

We are grateful to Dr. Jun Wen for her constructive suggestions in early thoughts of this work. This study was funded by the National Natural Science Foundation of China (grant nos. 31590821, 31622015, 31370261), Department of Science and Technology of Sichuan Province (grant 2015JQ0018) and Sichuan University. The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division.

Author contributions

K.M., M.R., J.L. designed research; K.M., M.R., Y.M., S.G., J.L., P.T., R.I.M., and P.M.H. performed research; M.R., K.M., Y.M. analyzed data; K.M., M.R., S.G., P.T., and P.M.H. procured specimens; K.M., M.R., R.I.M. wrote the manuscript, all authors revised the manuscript, and K.M., M.R., S.G., R.I.M. finalized the manuscript.

References

Adams R, Bartel J, Price R. 2009. A new genus, Hesperocyparis, for the cypresses of the Western Hemisphere (Cupressaceae). *Phytologia :* **91**(1): 160-185.

Adams RP. 2014. *Junipers of the World: The genus Juniperus*. Vancouver, B.C.: Trafford Publishing.

Barrie FR. 2011. Report of the General Committee: 11. *Taxon* **60**(4): 1211-1214.

Beiko RG, Doolittle WF, Charlebois RL. 2008. The impact of reticulate evolution on genome phylogeny. *Systematic Biology* **57**(6): 844-856.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15): 2114-2120.

Chamala S, Garcia N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, Barbazuk WB, Soltis DE, Soltis PS. 2015. Markerminer 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* **3**(4): 1400115.

646 **Christenhusz MJM, Reveal JL, Farjon A, Gardner MF, Mill RR, Chase MW. 2011.** A new classification and
647 linear sequence of extant gymnosperms. *Phytotaxa* **19**: 55-70.
648 **Dörken VM, Nimsch H, Jagel A. 2017.** Morphology, anatomy and morphogenesis of seed cones of *Cupressus*
649 *vietnamensis* (Cupressaceae) and the taxonomic and systematic implications. *Flora* **230**: 47-56.
650 **Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005.** Bayesian coalescent inference of past population
651 dynamics from molecular sequences. *Molecular Biology and Evolution* **22**(5): 1185-1192.
652 **Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe**
653 **GD, *et al*. 2008.** Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*
654 **452**(7188): 745-U745.
655 **Edwards SV. 2009.** Is a New and General Theory of Molecular Systematics Emerging? *Evolution* **63**(1): 1-19.
656 **Edwards SV, Liu L, Pearl DK. 2007.** High-resolution species trees without concatenation. *Proceedings of the*
657 *National Academy of Sciences USA* **104**(14): 5936-5941.
658 **Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012.** Ultraconserved
659 elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic*
660 *Biology* **61**(5): 717-726.
661 **Farjon A. 2005.** *A monograph of Cupressaceae and Sciadopitys*. Kew, U.K.: Royal Botanic Gardens, Kew.
662 **Felsenstein J**. **1978**. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic*
663 *Zoology* **27**: 401–410.
664 **Felstenstein, J. 1985.** Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.
665 **Gadek PA, Alpers DL, Heslewood MM, Quinn CJ. 2000.** Relationships within Cupressaceae sensu lato: a
666 combined morphological and molecular approach. *American Journal of Botany* **87**(7): 1044-1057.
667 **Heled J, Drummond AJ. 2010.** Bayesian inference of species trees from multilocus data. *Molecular Biology and*
668 *Evolution* **27**(3): 570-580.
669 **Hendy MD, Penny D. 1989**. A framework for quantitative study of evolutionary trees. *Systematic Zoology* **38**:
670 297–309.
671 **Huelsenbeck JP, Ronquist F. 2001.** MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8):
672 754-755.

673 **Jian SG, Soltis PS, Gitzendanner MA, Moore MJ, Li R, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE.**
674 **2008.** Resolving an ancient, rapid radiation in Saxifragales. *Systematic Biology* **57**(1): 38-57.
675 **Landis MJ, Matzke NJ, Moore BR. 2013.** Bayesian analysis of biogeography when the number of areas is
676 large. *Systematic Biology* 62: 789–804.
677 **Leache AD, Banbury BL, Linkem CW, de Oca ANM. 2016.** Phylogenomics of a rapid radiation: is chromosomal
678 evolution linked to increased diversification in north american spiny lizards (Genus *Sceloporus*)? *BMC*
679 *Evolutionary Biology* **16**: 63.
680 **Lee EK, Cibrian-Jaramillo A, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP,**
681 **Stevenson DW, McCombie WR, *et al*. 2011.** A functional phylogenomic view of the seed plants. *PLoS*
682 *Genetics* **7**(12): e1002411.
683 **Lemmon EM, Lemmon AR. 2013.** High-throughput genomic data in systematics and phylogenetics. *Annual*
684 *Review of Ecology, Evolution, and Systematics* **44**: 99-121.
685 **Li WZ, Godzik A. 2006.** Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide
686 sequences. *Bioinformatics* **22**(13): 1658-1659.
687 **Linder HP, Hardy CR, Rutschmann F. 2005.** Taxon sampling effects in molecular clock dating: An example from
688 the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**(3): 569-582.
689 **Little DP. 2006.** Evolution and circumscription of the true cypresses (Cupressaceae: *Cupressus*). *Systematic Botany*
690 **31**(3): 461-480.
691 **Little DP, Schwarzbach AE, Adams RP, Hsieh CF. 2004.** The circumscription and phylogenetic relateonships of
692 *Callitropsis* and the newly descriibed genus *Xanthocyparis* (Cupressaceae). *American Journal of Botany*
693 **91**(11): 1872-1881.
694 **Liu L, Yu LL, Pearl DK, Edwards SV. 2009a.** Estimating species phylogenies using coalescence times among
695 sequences. *Systematic Biology* **58**(5): 468-477.
696 **Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV. 2009b.** Coalescent methods for estimating phylogenetic trees.
697 *Molecular Phylogenetics and Evolution* **53**(1): 320-328.
698 **Liu L, Yu LL, Edwards SV. 2010.** A maximum pseudo-likelihood approach for estimating species trees under the
699 coalescent model. *BMC Evolutionary Biology* **10**: 302.

Liu L, Wu SY, Yu LL. 2015. Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution* **53**(5): 380-390.

Maddison MP. 1997. Gene trees in species trees. *Systematic Biology* **46**(3): 523–536.

Mao K, Hao G, Liu J, Adams RP, Milne RI. 2010. Diversification and biogeography of *Juniperus* (Cupressaceae): variable diversification rates and multiple intercontinental dispersals. *New Phytologist* **188**(1): 254-272.

Mao K, Milne RI, Zhang L, Peng Y, Liu J, Thomas P, Mill RR, Renner SS. 2012. Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences USA* **109**(20): 7793–7798.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10-12.

Matzke NJ. 2013. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Frontiers of Biogeography* **5**(4): 242-248.

Matzke NJ. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology* **63**(6): 951-970.

Mill RR, Farjon A. 2006. (1710) Proposal to conserve the name *Xanthocyparis* against *Callitropsis* Oerst. (Cupressaceae). *Taxon* **55**: 229-231.

Nylander J. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre Uppsala University.

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* **22**(1): 111-129.

Peng D, Wang XQ. 2008. Reticulate evolution in *Thuja* inferred from multiple gene sequences: Implications for the study of biogeographical disjunction between eastern Asia and North America. *Molecular Phylogenetics and Evolution* **47**(3): 1190-1202.

Pyron RA, Hendry CR, Chou VM, Lemmon EM, Lemmon AR, Burbrink FT. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution* **81**: 221-231.

Qu XJ, Jin JJ, Chaw SM, Li DZ, Yi TS. 2017. Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). *Scientific Reports* **7**: 41005.

Rambaut A, Drummond A 2009. Tracer v1.5, available from http://beast.bio.ed.ac.uk/Tracer.

Rannala B, Yang ZH. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**(4): 1645-1656.

Ree RH, Smith SA. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology* 57 (1): 4–14.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**(6960): 798-804.

Ronquist, F. 1997. Dispersal–vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic Biology* 46: 195–203.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12): 1572-1574.

Rothfels CJ, Li FW, Sigel EM, Huiet L, Larsson A, Burge DO, Ruhsam M, Deyholos M, Soltis DE, Stewart CN, Jr., *et al*. 2015. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *American Journal of Botany* **102**(7): 1089-1107.

Ruhsam M, Rai HS, Mathews S, Ross TG, Graham SW, Raubeson LA, Mei WB, Thomas PI, Gardner MF, Ennos RA, *et al*. 2015. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Molecular Ecology Resources* **15**(5): 1067-1078.

Rushforth K. 2007. Notes on the Cupressaceae in Vietnam. *Vietnam Journal of Biology* **29**(3): 32-39.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, *et al*. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). *Systematic Biology* **61**(2): 289-313.

Shaw TI, Ruan Z, Glenn TC, Liu L. 2013. STRAW: Species TRee Analysis Web server. *Nucleic Acids Research* **41**(W1): W238-W241.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9): 1312-1313.

Sun M, Soltis DE, Soltis PS, Zhu XY, Burleigh JG, Chen ZD. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Molecular Phylogenetics and Evolution* **83**: 156-166.

Swofford D 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0b10. Sunderland, MA, USA: Sinauer Associates.

Terry RG, Adams RP. 2015. A molecular re-examination of phylogenetic relationships among *Juniperus*, *Cupressus*, and the *Hesperocyparis-Callitropsis-Xanthocyparis* clades of Cupressaceae. *Phytologia* **97**(1): 67-75.

Wang XQ, Ran JH. 2014. Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution* **75**: 24-40.

Weisrock DW, Harmon LJ, Larson A. 2005. Resolving deep phylogenetic relationships in salamanders: Analyses of mitochondrial and nuclear genomic data. *Systematic Biology* **54**(5): 758-777.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, *et al*. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A* **111**(45): E4859-E4868.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends in Ecology & Evolution* **22**(5): 258-265.

Xiang Q, Li J. 2005. Derivation of *Xanthocyparis* and *Juniperus* from within *Cupressus*: evidence from sequences of nrDNA internal transcribed spacer region. *Harvard Papers in Botany* **9**(2): 375-382.

Xie YL, Wu GX, Tang JB, Luo RB, Patterson J, Liu SL, Huang WH, He GZ, Gu SC, Li SK, *et al*. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**(12): 1660-1666.

Yang ZY, Ran JH, Wang XQ. 2012. Three genome-based phylogeny of Cupressaceae s.l.: further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography. *Molecular Phylogenetics and Evolution* **64**(3): 452-470.

Yu Y, Harris AJ, Blair C, He XJ. 2015. RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution* **87**: 46-49.

Zeng LP, Zhang Q, Sun RR, Kong HZ, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* **5**: 4956.

Zuccon A, Zuccon D. 2014. MrEnt: an editor for publication-quality phylogenetic tree illustrations. *Molecular Ecology Resources* **14**(5): 1090-1094.

Tables

Table 1. A brief summary of alternative taxonomic treatments of *Juniperus*, *Cupressus*, *Hesperocyparis*, *Callitropsis* and *Xanthocyparis* since the description of *Xanthocyparis vietnamensis* in 2002. Underlined taxa are either monophyletic or monotypic. The abbreviations in brackets after common names are in accordance with Fig. 1.

(A)	Common names	Junipers	Old world cypresses (OWC)	New world cypresses (NWC)	Alaska cedar (A.)	Vietnamese golden cypress (V.)
(B)	This study; Adams *et al*. (2009); Mao *et al*. (2010, 2012)	*Juniperus*	*Cupressus* (s.s.)	*Hesperocyparis*	*Callitropsis* (s.s.)	*Xanthocyparis* (s.s.)
(C)	Farjon *et al*. (2002); Farjon (2005)	*Juniperus*	*Cupressus* sensu Farjon		*Xanthocyparis* s. l.	
(D)	Little *et al*. (2004)	N/A	*Cupressus* sensu Farjon		*Callitropsis* sensu Little (2004)	
(E)	Little (2006)	N/A	*Cupressus* (s.s.)	*Callitropsis* s.l.		
(F)	Christenhusz *et al*. (2011)	*Juniperus*	*Cupressus* s.l.			

797 **Table 2.** Accessions used for RNA extraction, transcriptome assembly and subsequent
798 phylogenetic analyses. (RBGE) and (SZ) refer to material collected from the wild held at the
799 Royal Botanic Garden Edinburgh and Sichuan University, respectively; (1kp) refers to
800 transcriptome data downloaded from the '1000 plant project' (http://www.onekp.com/samples/
801 list.php) with vouchers held at the University of British Columbia (UBC).
802

Species	Collecting number (Identifier)	Lat/Long	Country
Callitropsis nootkatensis	19941704B (RBGE)	49°24'N/123°11'W	Canada
Calocedrus decurrens	FRPM (1kp)	n/a	cultivated (UBC)
Cupressus duclouxiana	MSZ-49-01 (SZ)	27º00'N/100º14'E	China
C. funebris	Mao-CF (SZ)	n/a	cultivated (SZ)
C. gigantea	MSZ-24-03 (SZ)	29º00'N/93º14'E	China
C. sempervirens	19752308 (RBGE)	45°12'N/13°36'E	Croatia
Hesperocyparis arizonica	19921324*C (RBGE)	30°50'N/115°16'W	Mexico
H. bakeri	19851378*B (RBGE)	41°57'N/123°18'W	USA
H. macrocarpa	20090071 (RBGE)	36°31'N/121°56'W	USA
Juniperus drupacea	20100261 (RBGE)	37°55'N/36°34'E	Turkey
J. flaccida	19922158*C (RBGE)	25°17'N/100°26'W	Mexico
J. indica	19790193*A (RBGE)	27°13'N/88°02'E	India
J. microsperma	MSZ-11 (SZ)	29º37'N/96º20'E	China
J. oxycedrus	19921237A (RBGE)	37°54'N/2°52'W	Spain
J. phoenicea	19921233*A (RBGE)	37°54'N/2°52'W	Spain
J. procera	20080832*J (RBGE)	00°19'N/36°58'E	Kenya
J. scopulorum	20081601 (RBGE)	39°39'N/105°12'W	USA
Microbiota decussata	19881678*A (RBGE)	n/a	cultivated (RBGE)
M. decussata	XQSG (1kp)	n/a	cultivated (UBC)

Thuja plicata	VFYZ (1kp)	n/a	cultivated (UBC)
Xanthocyparis vietnamensis	20030523 (RBGE)	23°06'N/105°01'E	Vietnam

803

804

805 **Table 3.** Estimates for the divergence times for nodes within *Juniperus* and *Cupressus* (s.s.) and
806 the *Hesperocyparis-Callitropsis-Xanthocyparis* clade (HCX), based on the ptDNA data set of
807 Mao *et al*. (2010) using the constraint of the nuclear species tree topology from this study
808 (HCX(Cu,Ju) topology) or without any constraint (i.e. ptDNA tree topology, Ju(Cu,HCX)
809 topology) employing a relaxed molecular dating approach in BEAST.
810

Node No.	Description of Node:	HCX(Cu,Ju) topology	Ju(Cu,HCX) topology
1	Stem of the MRCA of the three clades	81.06 (70.45-90.40)	80.96 (71.07-89.75)
2	Crown of the MRCA of the three clades	59.80 (48.45-71.74)	59.44 (47.54-71.24)
3	Split between *Cupressus* and *Juniperus*	56.33 (45.30-67.95)	Equal to Node 2
4	Split between *Cupressus* and the HCX clade	Equal to Node 2	54.09 (41.29-67.03)
5	Crown of *Cupressus*	28.73 (11.65-42.15)	28.44 (16.57-42.03)
6	Crown of the HCX clade	37.45 (23.54-52.30)	36.41 (22.73-50.81)
7	Split between *Callitropsis* (s.s.) and *Hesperocyparis*	32.30 (19.40-45.86)	31.58 (19.13-44.84)
8	Crown of genus *Juniperus*	41.34 (33.90-49.45)	41.79 (33.91-50.80)
9	Split: sects. *Juniperus-Caryocedrus*	33.80 (19.68-45.58)	34.12 (20.28-46.87)
10	Crown of sect. *Juniperus*	17.20 (8.63-27.41)	17.16 (8.34-27.05)
11	Crown of sect. *Sabina*	36.50 (29.99-44.53)	36.80 (29.49-44.86)

811

812

Figure descriptions

Figure 1. Bayesian tree based on 73 concatenated nuclear genes (208.484 bp). Numbers or asterisks above branches are statistical support values for maximum parsimony/maximum likelihood/Bayesian inference analyses, respectively, with * denoting maximum support in all three analyses. Colour and grey scale bars to the right of the cladogram illustrate (A) common names of all included taxa (OWC = Old World cypresses; NWC = New World cypresses; A. = Alaska cedar; V. = Vietnamese golden cypress; *C.* = *Callitropsis*; *X.* = *Xanthocyparis*; sl/s.l. = sensu lato; sL = sensu Little (2004)), taxonomic treatments adopted (B) in this study, Adams *et al.* (2009) and Mao *et al.* (2010, 2012), (C) by Farjon *et al.* (2002) and Farjon (2005), (D) by Little *et al.* (2004), (E) by Little (2006) and (F) by Christenhusz *et al*. (2011). Scale bar indicates the estimated number of mutations per site.

Figure 2. Species tree generated using ASTRAL-II based on 73 nuclear genes (208.484 bp). Numbers or asterisks above branches are branch support values for MP-EST, STAR and ASTRAL-II analyses, respectively, with asterisks denoting maximum support in all three analyses. ASTRAL-II measures branch lengths in coalescent units (scale bar shown corresponds to two coalescent unit) for internal branches and NOT terminal branches (branch lengths of terminal branches are therefore arbitrary and meaningless). The pie chart shows respective quartet support for the main topology, the first and the second alternative topology (as shown in the inset). Note that in the inset, the tree formulas in parentheses were presented in the sense of an unrooted quadripartition (four-taxon) tree, where the central piece of the tree is an internode branch between the two pairs of partitions. The first case shown ((*Cupressus*, *Juniperus*),(HCX, outgroups)) is consistent with HCX being sister to (*Cupressus*, *Juniperus*) in an outgroup-rooted tree.

Figure 3. Neighbour-Net networks based on 73 concatenated nuclear genes (208.484 bp) using SplitsTree. (A) A Neighbour-Net network with bootstrap support values, and (B) a consensus Neighbour-Net network using a 95% threshold based on 1,000 bootstrap replicates. In (A), numbers next to "branches" are bootstrap support values. Note that in both (A) and (B) the branches lead to the three outgroup taxa are truncated so as to show more details of ingroup relationships.

Figure 4. Evolutionary divergence timescale of Cupressoideae based on the ptDNA data set (86 accessions from Mao *et al*., 2010) with the imposed constraint of the nuclear species tree (HCX(Cu,Ju)) topology from this study using BEAST. Blue bars represent 95% HPD (highest posterior density) for each node, and white triangular with black outline represent compressed clades. Letters in black circles represent fossil calibration points (corresponding to those in Table 1 in Mao *et al*., 2010), and numbers in black squares indicate numbers for nodes of interest (see Table 3). 'HCX' stands for the *Hesperocyparis-Callitropsis-Xanthocyparis* clade.

Figure 5. Ancestral area reconstruction based on BEAST trees constrained using nuclear species tree (HCX(Cu,Ju)) topology and the DIVALIKE+J model in BioGeoBEARS, as implemented in RASP 4.0. The inset shows the division of the distribution area of the five genera into four operational areas (North America, Africa, Asia and Europe). The pie at each node represents the reconstructed ancestral area; different colours of circular sector in a pie represent the relative probabilities of different ancestral areas at a node.

Supporting Information

Supplementary Table S1. Detailed information for the 73 single/low copy nuclear genes used in phylogenetic analyses. The 'missing taxa' column highlights which taxa were not included due to missing data for a particular gene. Missing taxa in bold are from the ingroup; a list of acronyms and corresponding species names is given below the table.

Supplementary Table S2. NCBI GenBank accession numbers for the 73 genes of 20 taxa that were used in phylogenetic analyses.

Supplementary Table S3. A summary of the topology and the maximum likelihood (ML) bootstrap support value (MLBS) for each of the 73 nuclear genes used to construct the species tree. Different topologies showing the sister relationship between the three major clades, *Juniperus*, *Cupressus* and HCX, are shown. 'Other topology' refers to a situation where two or all of the three groups are polyphyletic and the topology does not fall into any of the first three categories. 'Polyphyletic group(s)' refers to a situation where one or more taxa are polyphyletic. 'MLBS value' refers to the ML bootstrap support value for the sister relationship between the indicated clade.

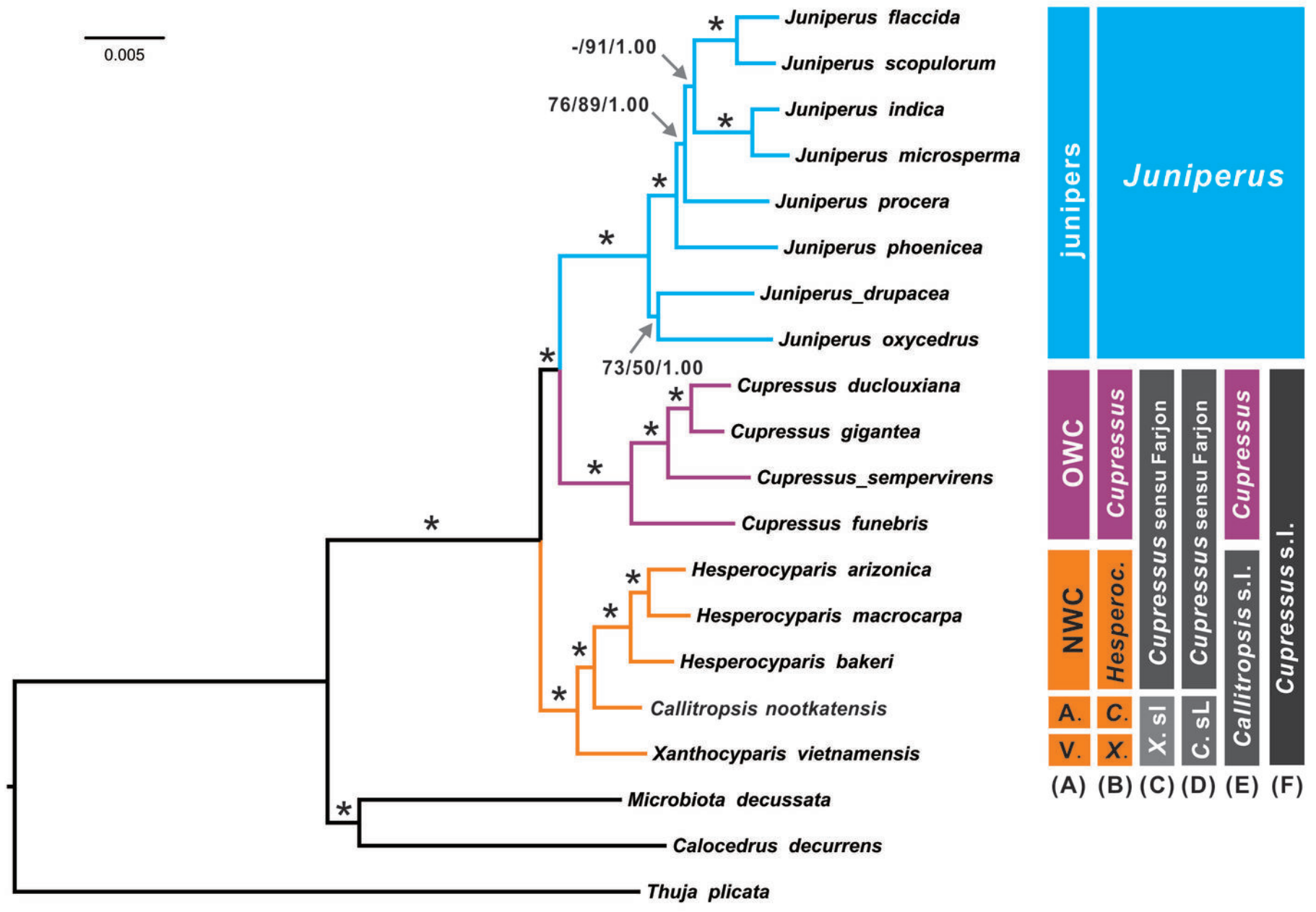
Supplementary Figure S1. Maximum parsimony (MP) tree based on 73 concatenated nuclear genes (208.484 bp). MP analysis resulted in one tree of 31,272 steps with CI = 0.86 and RI = 0.82. Scale bar indicates number of nucleotide changes. Numbers above or below branches are bootstrap support values based on 1000 bootstrap replicates.

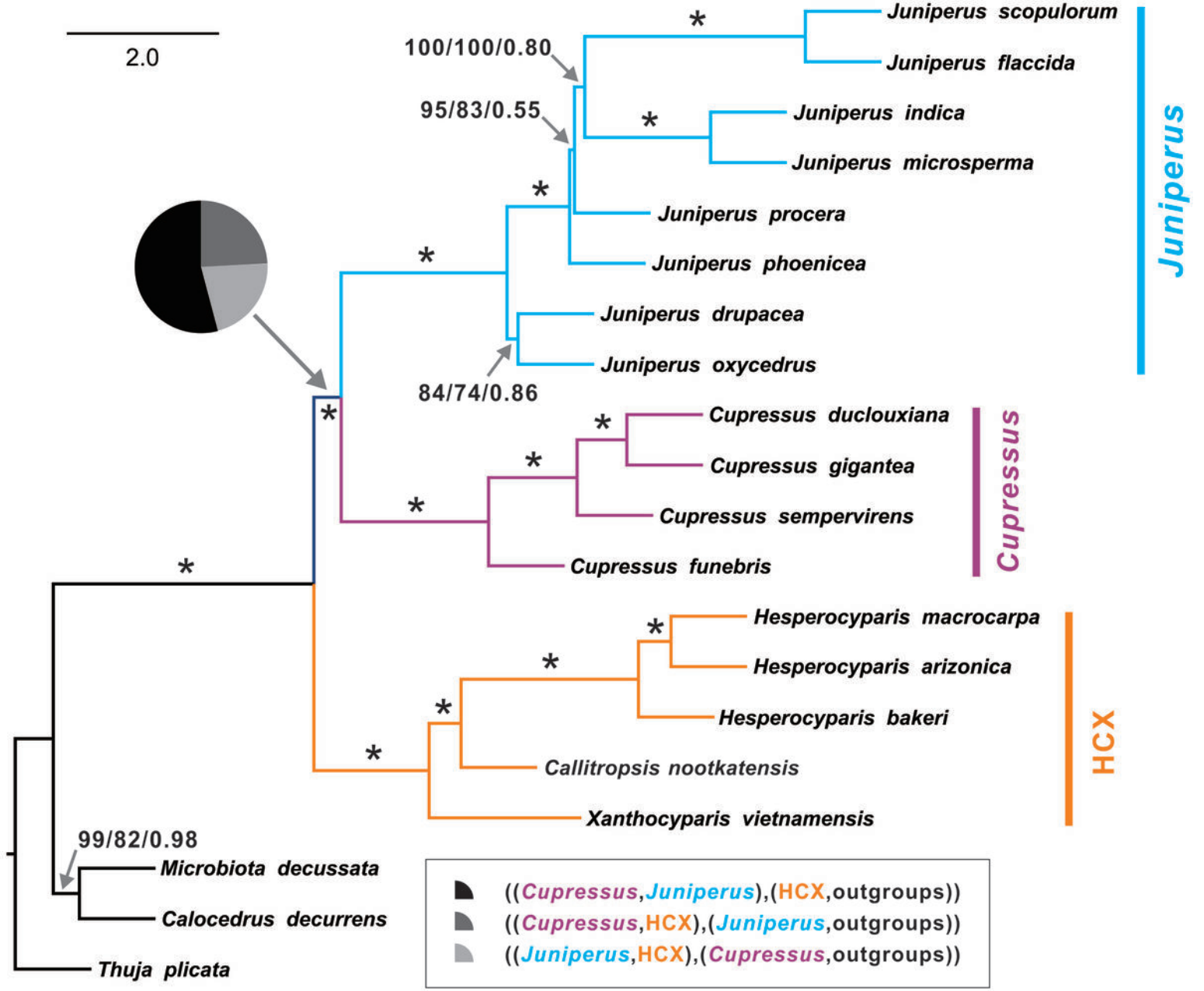
Supplementary Figure S2. Phylogenetic relationships and posterior probability of major clades in Cupressoideae that were derived from BEAST using nine ptDNA fragments (86 accessions

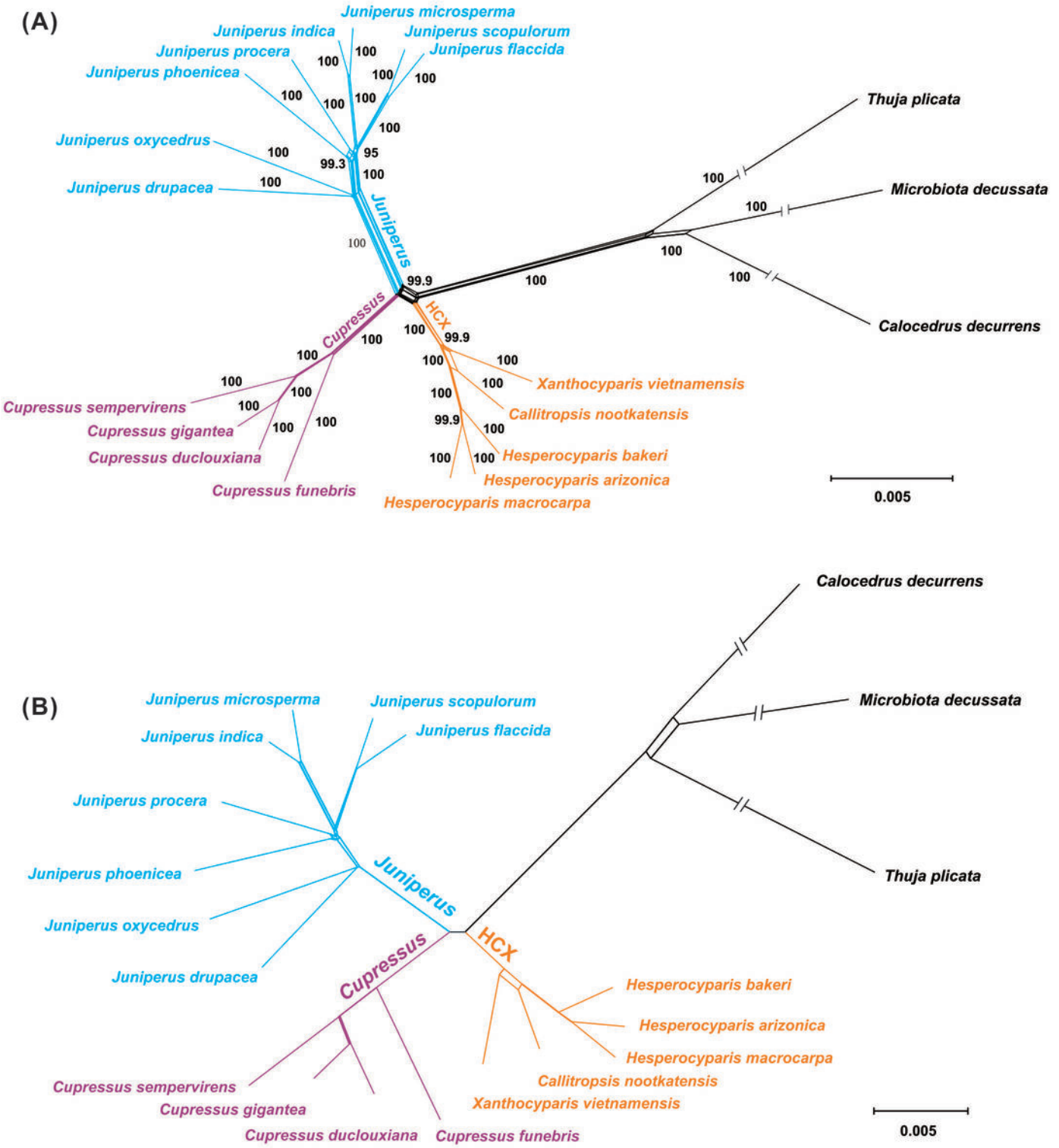
from Mao *et al*., 2010) and eight fossil calibrations (a) without any constraint (i.e. ptDNA tree, Ju(Cu,HCX) topology) and (b) with a constraint using the nuclear species tree (HCX(Cu,Ju)) topology concerning the relationship of *Cupressus*, *Juniperus* and the HCX clade (only) that was obtained in this study (see Fig. 1). Triangle tips represent clades that comprising more than two species; for full list of species see Figs 5 and S4, respectively. Posterior probabilities of clades shown to the right or upper-left of each node.

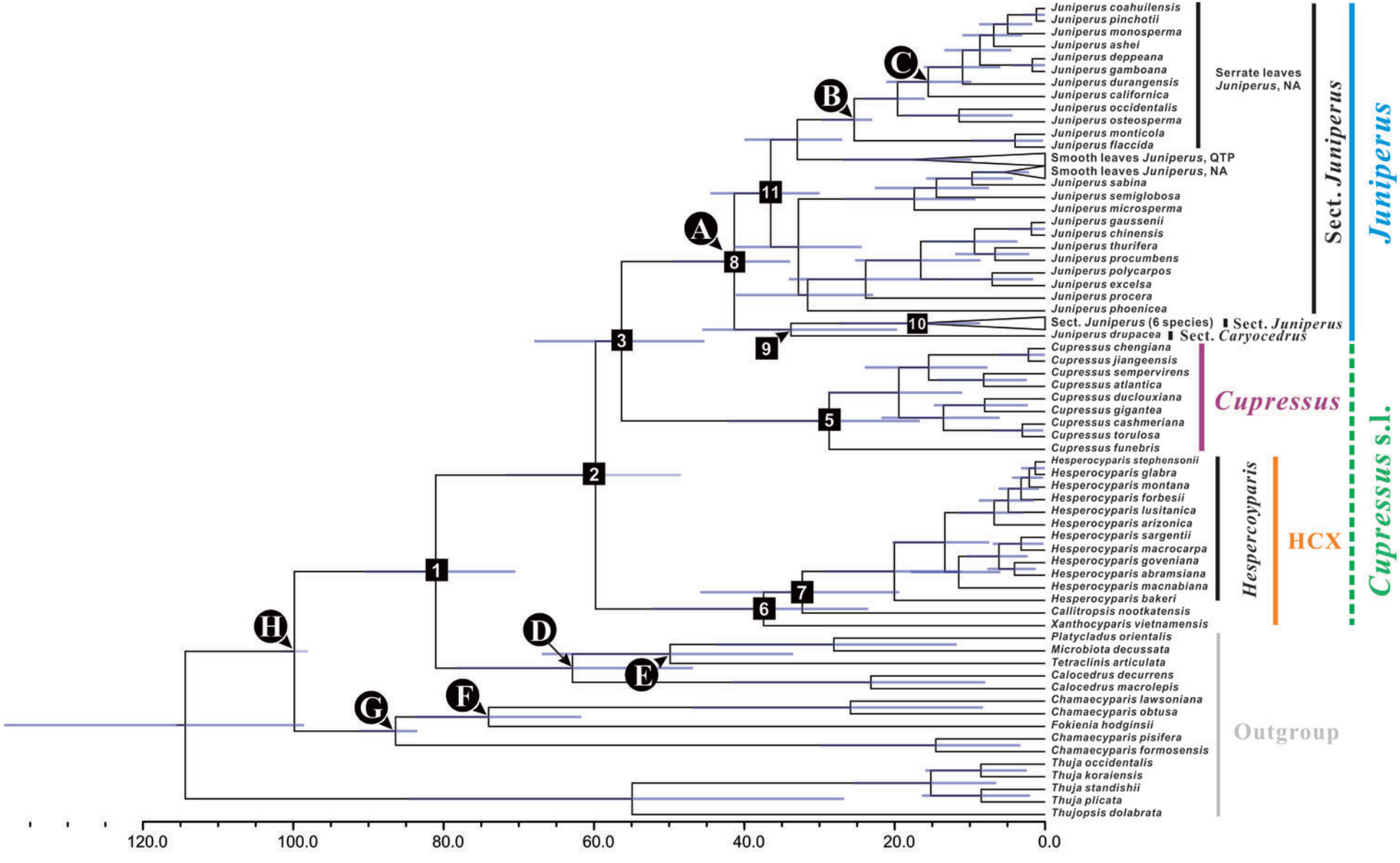
Supplementary Figure S3. Maximum likelihood bootstrap tree for each of the 73 nuclear genes as constructed in RaxML based on 1000 bootstrap replicates. Gene names are shown in the top left of each subfigures. Numbers close to each node represent the bootstrap support value for the branch lead to a node.

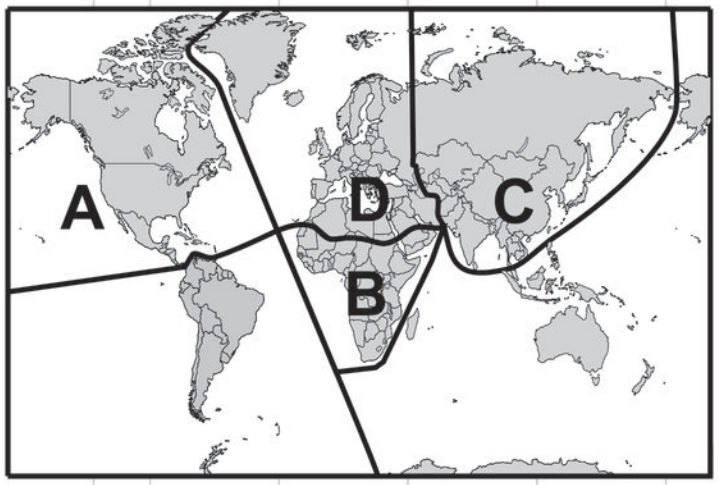
Supplementary Figure S4. Ancestral area reconstruction based on ptDNA tree (Ju(Cu,HCX)) topology and the DIVALIKE+J model in BioGeoBEARS as implemented in RASP 4.0. The inset shows the division of the distribution area of the five genera into four operational areas, North America, Africa, Asia and Europe. The pie at each node represents the reconstructed ancestral area, different colours of circular sectors in a pie represent the relative probabilities of different ancestral areas at a node.


0.005
-/91/1.00
76/89/1.00
73/50/1.00
Juniperus flaccida
Juniperus scopulorum
Juniperus indica
Juniperus microsperma
Juniperus procera
Juniperus phoenicea
Juniperus_drupacea
Juniperus oxycedrus
Cupressus duclouxiana
Cupressus gigantea
Cupressus_sempervirens
Cupressus funebris
Hesperocyparis arizonica
Hesperocyparis macrocarpa
Hesperocyparis bakeri
Callitropsis nootkatensis
Xanthocyparis vietnamensis
Microbiota decussata
Calocedrus decurrens
Thuja plicata
junipers
Juniperus
OWC
NWC
Cupressus
Hesperoc.
Cupressus sensu Farjon
Cupressus sensu Farjon
Cupressus
Callitropsis s.l.
Cupressus s.l.
A.
V.
C.
X.
X. sl
C. sL
(A)
(B)
(C)
(D)
(E)
(F)


2.0
100/100/0.80
95/83/0.55
84/74/0.86
99/82/0.98
Juniperus scopulorum
Juniperus flaccida
Juniperus indica
Juniperus microsperma
Juniperus procera
Juniperus phoenicea
Juniperus drupacea
Juniperus oxycedrus
Juniperus
Cupressus duclouxiana
Cupressus gigantea
Cupressus sempervirens
Cupressus funebris
Cupressus
Hesperocyparis macrocarpa
Hesperocyparis arizonica
Hesperocyparis bakeri
Callitropsis nootkatensis
Xanthocyparis vietnamensis
HCX
Microbiota decussata
Calocedrus decurrens
Thuja plicata
((Cupressus,Juniperus),(HCX,outgroups))
((Cupressus,HCX),(Juniperus,outgroups))
((Juniperus,HCX),(Cupressus,outgroups))


(A)
Juniperus microsperma
Juniperus indica
Juniperus scopulorum
Juniperus procera
Juniperus flaccida
Juniperus phoenicea
Juniperus oxycedrus
Juniperus drupacea
Juniperus
Thuja plicata
Microbiota decussata
Calocedrus decurrens
Cupressus
HCX
Xanthocyparis vietnamensis
Callitropsis nootkatensis
Hesperocyparis bakeri
Hesperocyparis arizonica
Hesperocyparis macrocarpa
Cupressus sempervirens
Cupressus gigantea
Cupressus duclouxiana
Cupressus funebris
0.005
(B)
Juniperus microsperma
Juniperus scopulorum
Juniperus indica
Juniperus flaccida
Juniperus procera
Juniperus phoenicea
Juniperus oxycedrus
Juniperus drupacea
Juniperus
Calocedrus decurrens
Microbiota decussata
Thuja plicata
HCX
Cupressus
Hesperocyparis bakeri
Hesperocyparis arizonica
Hesperocyparis macrocarpa
Callitropsis nootkatensis
Xanthocyparis vietnamensis
Cupressus sempervirens
Cupressus gigantea
Cupressus duclouxiana
Cupressus funebris
0.005


Juniperus coahuilensis
Juniperus pinchotii
Juniperus monosperma
Juniperus ashei
Juniperus deppeana
Juniperus gamboana
Juniperus durangensis
Juniperus californica
Juniperus occidentalis
Juniperus osteosperma
Juniperus monticola
Juniperus flaccida
Smooth leaves Juniperus, QTP
Smooth leaves Juniperus, NA
Juniperus sabina
Juniperus semiglobosa
Juniperus microsperma
Juniperus gaussenii
Juniperus chinensis
Juniperus thurifera
Juniperus procumbens
Juniperus polycarpos
Juniperus excelsa
Juniperus procera
Juniperus phoenicea
Sect. Juniperus (6 species)
Juniperus drupacea
Cupressus chengiana
Cupressus jiangeensis
Cupressus sempervirens
Cupressus atlantica
Cupressus duclouxiana
Cupressus gigantea
Cupressus cashmeriana
Cupressus torulosa
Cupressus funebris
Hesperocyparis stephensonii
Hesperocyparis glabra
Hesperocyparis montana
Hesperocyparis forbesii
Hesperocyparis lusitanica
Hesperocyparis arizonica
Hesperocyparis sargentii
Hesperocyparis macrocarpa
Hesperocyparis goveniana
Hesperocyparis abramsiana
Hesperocyparis macnabiana
Hesperocyparis bakeri
Callitropsis nootkatensis
Xanthocyparis vietnamensis
Platycladus orientalis
Microbiota decussata
Tetraclinis articulata
Calocedrus decurrens
Calocedrus macrolepis
Chamaecyparis lawsoniana
Chamaecyparis obtusa
Fokienia hodginsii
Chamaecyparis pisifera
Chamaecyparis formosensis
Thuja occidentalis
Thuja koraiensis
Thuja standishii
Thuja plicata
Thujopsis dolabrata
Serrate leaves Juniperus, NA
Sect. Juniperus
Juniperus
Sect. Juniperus
Sect. Caryocedrus
Cupressus
Hespercoyparis
HCX
Cupressus s.l.
Outgroup
A
B
C
D
E
F
G
H
1
2
3
5
6
7
8
9
10
11
120.0
100.0
80.0
60.0
40.0
20.0
0.0



LEGEND

- * Ambiguous
- A North America
- AC North America + Asia
- B Africa
- BC Africa + Asia
- C Asia
- CD Asia+Europe
- D Europe

(A) *Juniperus coahuilensis*
(A) *Juniperus pinchotii*
(A) *Juniperus monosperma*
(A) *Juniperus ashei*
(A) *Juniperus deppeana*
(A) *Juniperus gamboana*
(A) *Juniperus durangensis*
(A) *Juniperus californica*
(A) *Juniperus occidentalis*
(A) *Juniperus osteosperma*
(A) *Juniperus monticola*
(A) *Juniperus flaccida*
(C) *Juniperus saltuaria*
(C) *Juniperus przewalskii*
(C) *Juniperus pingii*
(C) *Juniperus recurva*
(C) *Juniperus convallium*
(C) *Juniperus kamarovii*
(C) *Juniperus tibetica*
(C) *Juniperus indica*
(C) *Juniperus squamata*
(C) *Juniperus pseudosabina*
(A) *Juniperus bermudiana*
(A) *Juniperus virginiana*
(A) *Juniperus horizeontalis*
(A) *Juniperus scopulorum*
(A) *Juniperus saxicola*
(A) *Juniperus lucayana*
(CD) *Juniperus sabina*
(C) *Juniperus semiglobosa*
(C) *Juniperus microsperma*
(C) *Juniperus gaussenii*
(C) *Juniperus chinensis*
(D) *Juniperus thurifera*
(C) *Juniperus procumbens*
(C) *Juniperus polycarpos*
(D) *Juniperus excelsa*
(B) *Juniperus procera*
(D) *Juniperus phoenicea*
(A) *Juniperus communis* var. *depressa*
(D) *Juniperus communis* var. *communis*
(C) *Juniperus communis* var. *saxillis*
(C) *Juniperus taxifolia*
(C) *Juniperus rigida*
(C) *Juniperus formosana*
(D) *Juniperus oxycedrus*
(D) *Juniperus deltoides*
(D) *Juniperus drupacea*
(C) *Cupressus chengiana*
(C) *Cupressus jiangeensis*
(D) *Cupressus sempervirens*
(D) *Cupressus atlantica*
(C) *Cupressus duclouxiana*
(C) *Cupressus gigantea*
(C) *Cupressus cashmeriana*
(C) *Cupressus torulosa*
(C) *Cupressus funebris*
(A) *Hesperocyparis stephensonii*
(A) *Hesperocyparis glabra*
(A) *Hesperocyparis montana*
(A) *Hesperocyparis forbesii*
(A) *Hesperocyparis lusitanica*
(A) *Hesperocyparis arizonica*
(A) *Hesperocyparis sargentii*
(A) *Hesperocyparis macrocarpa*
(A) *Hesperocyparis goveniana*
(A) *Hesperocyparis abramsiana*
(A) *Hesperocyparis macnabiana*
(A) *Hesperocyparis bakeri*
(A) *Callitropsis nootkatensis*
(C) *Xanthocyparis vietnamensis*
(C) *Platycladus orientalis*
(C) *Microbiota decussata*
(D) *Tetraclinis articulata*
(A) *Calocedrus decurrens*
(C) *Calocedrus macrolepis*
(A) *Chamaecyparis lawsoniana*
(C) *Chamaecyparis obtusa*
(C) *Fokienia hodginsii*
(C) *Chamaecyparis pisifera*
(C) *Chamaecyparis formosensis*
(A) *Thuja occidentalis*
(C) *Thuja koraiensis*
(C) *Thuja standishii*
(A) *Thuja plicata*
(C) *Thujopsis dolabrata*

Sect. *Sabina*
Sect. *Juniperus*
Juniperus
Sect. *Caryocedrus*
Cupressus
Hesperocyparis
HCX
Cal.
Xan.
Cupressus s.l.
Outgroup

Early Cretaceous | Late Cretaceous | Paleo. | Eocene | Oligo. | Miocene | P | Q
Mesozoic | Cenozoic
110 100 90 80 70 60 50 40 30 20 10 0 (Mya)